Mellanox TECHNOLOGIES

Connect. Accelerate. Outperform.™

Mellanox NATIVE ESX Driver for VMware vSphere 6.0 User Manual

Rev 3.2.0.15

www.mellanox.com

NOTE:
THIS HARDWARE, SOFTWARE OR TEST SUITE PRODUCT (“PRODUCT(S)”) AND ITS RELATED DOCUMENTATION ARE PROVIDED BY MELLANOX TECHNOLOGIES “AS-IS” WITH ALL FAULTS OF ANY KIND AND SOLELY FOR THE PURPOSE OF AIDING THE CUSTOMER IN TESTING APPLICATIONS THAT USE THE PRODUCTS IN DESIGNATED SOLUTIONS. THE CUSTOMER'S MANUFACTURING TEST ENVIRONMENT HAS NOT MET THE STANDARDS SET BY MELLANOX TECHNOLOGIES TO FULLY QUALIFY THE PRODUCT(S) AND/OR THE SYSTEM USING IT. THEREFORE, MELLANOX TECHNOLOGIES CANNOT AND DOES NOT GUARANTEE OR WARRANT THAT THE PRODUCTS WILL OPERATE WITH THE HIGHEST QUALITY. ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT ARE DISCLAIMED. IN NO EVENT SHALL MELLANOX BE LIABLE TO CUSTOMER OR ANY THIRD PARTIES FOR ANY DIRECT, INDIRECT, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES OF ANY KIND (INCLUDING, BUT NOT LIMITED TO, PAYMENT FOR PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY FROM THE USE OF THE PRODUCT(S) AND RELATED DOCUMENTATION EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

Mellanox Technologies
350 Oakmead Parkway Suite 100
Sunnyvale, CA 94085
U.S.A.
www.mellanox.com
Tel: (408) 970-3400
Fax: (408) 970-3403

Table of Contents

List of Tables

Document Revision History

Table 1 - Document Revision History

Release	Date	Description
3.2.0.15	November, 2015	Updated the following sections: • Section 2.2, “Installing Mellanox NATIVE ESX Driver for VMware vSphere”, on page 12 • Section 2.3, “Removing Mellanox OFED Driver”, on page 13 • Section 2.4, “Loading/Unloading Driver Kernel Modules”, on page 13
3.2.0	November, 2015	Updated section Section 3.1, “VXLAN Hardware Offload”, on page 15
	September, 2015	Initial release of this MLNX-NATIVE-ESX version

About this Manual

This preface provides general information concerning the scope and organization of this User's Manual.

Intended Audience

This manual is intended for system administrators responsible for the installation, configuration, management and maintenance of the software and hardware of VPI (in Ethernet mode), and Ethernet adapter cards. It is also intended for application developers.

Common Abbreviations and Acronyms

Table 2 - Abbreviations and Acronyms (Sheet 1 of 2)

Abbreviation / Acronym	Whole Word / Description
B	(Capital) 'B' is used to indicate size in bytes or multiples of bytes (e.g., 1KB = 1024 bytes, and 1MB = 1048576 bytes)
b	(Small) 'b' is used to indicate size in bits or multiples of bits (e.g., 1Kb = 1024 bits)
FW	Firmware
HCA	Host Channel Adapter
HW	Hardware
LSB	Least significant *byte*
lsb	Least significant *bit*
MSB	Most significant *byte*
msb	Most significant *bit*
NIC	Network Interface Card
SW	Software
VPI	Virtual Protocol Interconnect
PR	Path Record
RDS	Reliable Datagram Sockets
SDP	Sockets Direct Protocol
SL	Service Level
MPI	Message Passing Interface
QoS	Quality of Service
ULP	Upper Level Protocol

Table 2 - Abbreviations and Acronyms (Sheet 2 of 2)

Abbreviation / Acronym	Whole Word / Description
vHBA	Virtual SCSI Host Bus adapter
uDAPL	User Direct Access Programming Library

Related Documentation

Table 3 - Reference Documents

Document Name	Description
IEEE Std 802.3ae™-2002 (Amendment to IEEE Std 802.3-2002) Document # PDF: SS94996	Part 3: Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications Amendment: Media Access Control (MAC) Parameters, Physical Layers, and Management Parameters for 10 Gb/s Operation
Firmware Release Notes for Mellanox adapter devices	See the Release Notes PDF file relevant to your adapter device. For further information please refer to the Mellanox website. www.mellanox.com -> Support -> Firmware Download
MFT User Manual	Mellanox Firmware Tools User's Manual. For further information please refer to the Mellanox website. www.mellanox.com -> Products -> InfiniBand/VPI Driver -> Firmware Tools
MFT Release Notes	Release Notes for the Mellanox Firmware Tools. For further information please refer to the Mellanox website. www.mellanox.com -> Products -> InfiniBand/VPI Driver -> Firmware Tools
VMware vSphere 6.0 Documentation Center	VMware website

1 Introduction

Mellanox NATIVE ESX is a software stack which operates across Mellanox's ConnectX®-3 and ConnectX®-3 Pro network adapter solutions supporting up to 40Gb/s Ethernet (ETH) and 2.5 or 5.0 GT/s PCI Express 2.0 and 3.0 uplinks to servers.

The following sub-sections briefly describe the various components of the Mellanox NATIVE ESX stack.

1.1 nmlx4 Driver

`nmlx4` is the low level driver implementation for the ConnectX®-3 and ConnectX®-3 Pro adapters designed by Mellanox Technologies. ConnectX® family adapters can operate as an InfiniBand adapter, or as an Ethernet NIC. The NATIVE ESX driver supports Ethernet NIC configurations. To accommodate the supported configurations, the driver is split into the following modules:

nmlx4_core

Handles low-level functions like device initialization and firmware commands processing. Also controls resource allocation.

nmlx4_en

A 10/40GigE driver that handles Ethernet specific functions and plugs into the ESX uplink layer

1.2 Ethernet (EN) Management Interface

EN Management Interface provides user space application means to access EN driver data (similar to ethtool interface in Linux).

The kernel space management interface layer is implemented as part of the EN driver and released as part of nmlx4_en.

The user space management interface access layer provides a set of accessors functions to driver objects and it is provided as a development source package tarball (H and C files). The user should include the H file and compile and link against the C file.

1.3 Mellanox Firmware Tools

The Mellanox Firmware Tools (MFT) package is a set of firmware management tools for a single node. MFT can be used for:

- Generating a standard or customized Mellanox firmware image
- Burning a firmware image to a single node

MFT includes the following tools:

- flint: burns a firmware binary image or an expansion ROM image to the Flash device of a Mellanox network adapter/bridge/switch device. It includes query functions to the burnt firmware image and to the binary image file.
- Debug utilities: A set of debug utilities (e.g., itrace, mlxdump, mstdump, mlxmcg, wqdump, mcra, i2c, mget_temp, and pckt_drop)

For additional details, please refer to the MFT User's Manual
www.mellanox.com -> Products -> InfiniBand/VPI Driver -> Firmware Tools.

1.4 Mellanox NATIVE ESX Package

1.4.1 Software Components

MLNX-NATIVE-ESX contains the following software components:

- Mellanox Host Channel Adapter Drivers
  - nmlx4 which is split into multiple modules:
    - nmlx4_core (low-level helper)
    - nmlx4_en (Ethernet)
- EN Management Interface development package tarball

1.5 Module Parameters

1.5.1 nmlx4 Module Parameters

To set **nmlx4** parameters:

```
esxcli system module parameters set -m nmlx4_core -p '<parameter>=<value>'
```

and/or

```
esxcli system module parameters set -m nmlx4_en -p '<parameter>=<value>'
```

To show all parameters which were set until now:

```
esxcli system module parameters list -m  <module name>
```

Parameters which are not set by the user, remain on default value.

The following sections list the available `nmlx4` parameters.

1.5.1.1 nmlx4_core Parameters

Table 1 - nmlx4_core Parameters

Name	Description	Values
`enable_64b_cqe_eqe`	Enables 64 byte CQEs/EQEs when the firmware supports this.	• 1 - enabled • 0 - disabled Default: 0
`enable_dmfs`	Enables Device Managed Flow Steering	• 1 - enabled • 0 - disabled Default: 1
`enable_qos`	Enables Quality of Service support in the HCA	• 1 - enabled • 0 - disabled Default: 0
`enable_rocev2`[1]	Enables RoCEv2 mode for all devices	• 1 - enabled • 0 - disabled Default: 0
`enable_vxlan_offloads`	Enables VXLAN offloads when supported by NIC	• 1 - enabled • 0 - disabled Default: 1
`log_mtts_per_seg`	Log2 number of MTT entries per segment	1-7 Default: 3
`log_num_mgm_entry_size`	Log2 MGM entry size, that defines the number of QPs per MCG, for example: value 10 results in 248 QP per MGM entry	9-12 Default: 12
`msi_x`	Enables MSI-X	• 1 - enabled • 0 - disabled Default: 1
`mst_recovery`	Enables recovery mode (only NMST module is loaded)	• 1 - enabled • 0 - disabled Default: 0
`rocev2_udp_port`[1]	Destination port for RoCEv2	1-65535 for RoCEv2 Default: 4791

[1] The following module parameter is not supported in this version.

1.5.1.2 nmlx4_en Parameters

Table 2 - nmlx4_en Parameters

Name	Description	Value
`num_rings_per_rss_queue`	Enables RSS	• 2-4 • 0 - disabled Default: 0 When this value is != 0, RSS is enabled with 1 RSS Queue that manages `num_rings_per_rss_queue` Rx Rings **Note:** The value must be a power of 2
`pfcrx`	Priority based Flow Control policy on RX.	0-255 Default: 0 It is a 8 bits bit mask, each bit indicates priority [0-7]. • 1 respects incoming pause frames on the specified priority. • 0 - ignore incoming pause frames on the specified priority.
`pfctx`	Priority based Flow Control policy on TX.	0-255 Default: 0 It is a 8 bits bit mask, each bit indicates priority [0-7]. • 1 - generates pause frames according to the RX buffer threshold on the specified priority. • 0 - never generates pause frames on the specified priority.
`vlan_filter`	Enables VLAN filter	• 1 - enabled • 0 - disabled Default: 0

1.6 Device Capabilities

Normally, an application needs to query the device capabilities before attempting to create a resource. It is essential for the application to be able to operate over different devices with different capabilities.

Specifically, when creating a QP, the user needs to specify the maximum number of outstanding work requests that the QP supports. This value should not exceed the queried capabilities. However, even when you specify a number that does not exceed the queried capability, the verbs can still fail since some other factors such as the number of scatter/gather entries requested, or the size of the inline data required, affect the maximum possible work requests. Hence an application should try to decrease this size (halving is a good new value) and retry until it succeeds.

2 Installation

This chapter describes how to install and test the Mellanox NATIVE ESX package on a single host machine with Mellanox Ethernet adapter hardware installed.

2.1 Hardware and Software Requirements

Table 3 - Software and Hardware Requirements

Requirements	Description
Platforms	A server platform with an adapter card based on one of the following Mellanox Technologies’ HCA devices: • MT27508 ConnectX®-3 (VPI, EN) (firmware: fw-ConnectX3) • MT4103 ConnectX®-3 Pro (VPI, EN) (firmware: fw-ConnectX3Pro)
Device ID	For the latest list of device IDs, please visit Mellanox website.
Operating System	ESXi 2015 operating system.
Installer Privileges	The installation requires administrator privileges on the target machine.

2.2 Installing Mellanox NATIVE ESX Driver for VMware vSphere

Please uninstall any previous Mellanox driver packages prior to installing the new version.

➢ ***To install the driver:***

1. Log into the ESXi server with root permissions.
2. Install the driver.

```
#> esxcli software vib install -d <path>/<bundle_file>
```

Example:

```
#> esxcli software vib install -d <path>/MLNX-NATIVE-ESX-ConnectX-3-3.2.0.15-10EM-
600.0.0.2768847.zip
```

3. Reboot the machine.
4. Verify the driver was installed successfully.

```
# esxcli software vib list | grep mlx
nmlx4-core                  3.2.0.15-1OEM.600.0.0.2768847      MEL      PartnerSupported  2015-11-15
nmlx4-en                    3.2.0.15-1OEM.600.0.0.2768847      MEL      PartnerSupported  2015-11-15
nmlx4-rdma                  3.2.0.15-1OEM.600.0.0.2768847     MEL      PartnerSupported  2015-11-15
```

After the installation process, all kernel modules are loaded automatically upon boot.

2.3 Removing Mellanox OFED Driver

Please unload the driver before removing it.

➢ ***To remove all the drivers:***

1. Log into the ESXi server with root permissions.
2. List the existing NATIVE ESX driver modules. (see Step 5 in Section 2.2, on page 12)
3. Remove each module.

```
#> esxcli software vib remove -n nmlx4-rdma
#> esxcli software vib remove -n nmlx4-en
#> esxcli software vib remove -n nmlx4-core
```

To remove the modules, the command must be run in the same order as shown in the example above.

4. Reboot the server.

2.4 Loading/Unloading Driver Kernel Modules

➢ ***To unload the driver:***

```
esxcfg-module -u nmlx4_rdma
esxcfg-module -u nmlx4_en
esxcfg-module -u nmlx4_core
```

➢ ***To load the driver:***

```
/etc/init.d/sfcbd-watchdog stop
esxcfg-module nmlx4_core
esxcfg-module nmlx4_en
esxcfg-module nmlx4_rdma
/etc/init.d/sfcbd-watchdog start
kill -POLL $(cat /var/run/vmware/vmkdevmgr.pid)
```

➢ ***To restart the driver:***

```
/etc/init.d/sfcbd-watchdog stop
esxcfg-module -u nmlx4_rdma
esxcfg-module -u nmlx4_en
esxcfg-module -u nmlx4_core
esxcfg-module nmlx4_core
esxcfg-module nmlx4_en
esxcfg-module nmlx4_rdma
/etc/init.d/sfcbd-watchdog start
kill -POLL $(cat /var/run/vmware/vmkdevmgr.pid)
```

2.5 Firmware Programming

1. Download the VMware bootable binary images v3.8.0 from the Mellanox Firmware Tools (MFT) site.
   - **File:** mft-3.8.0.56-10EM-600.0.0.2295424.x86_64.vib

     MD5SUM: 083baec399de55a181f5b26613ae0829
   - **File:** nmst-3.8.0.56-1OEM.600.0.0.2295424.x86_64.vib

     MD5SUM: 0426a9ab6e759ad44942d5061a6e9cfe
2. Install the image according to the steps described in the MFT User Manual.

> The following procedure requires custom boot image downloading, mounting and booting from a USB device.

3 Features Overview and Configuration

3.1 VXLAN Hardware Offload

VXLAN hardware offload enables the traditional offloads to be performed on the encapsulated traffic. With ConnectX®-3 Pro, data center operators can decouple the overlay network layer from the physical NIC performance, thus achieving native performance in the new network architecture.

3.1.1 Configuring VXLAN Hardware Offload

VXLAN hardware offload includes:

- TX: Calculates the Inner L3/L4 and the Outer L3 checksum
- RX:
  - Checks the Inner L3/L4 and the Outer L3 checksum
  - Maps the VXLAN traffic to an RX queue according to:
    - Inner destination MAC address
    - Outer destination MAC address
    - VXLAN ID

VXLAN hardware offload is enabled by default. However, if it was disable and you want to re-enable it, enable the `nmlx4_core` module parameters `"enable_vxlan_offloads"` and `"enable_dmfs"` (setting the parameters to 1).

➢ ***To enable VXLAN hardware offload:***

```
esxcli system module parameters set -m nmlx4_core -p 'enable_vxlan_offloads=1'
esxcli system module parameters set -m nmlx4_core -p 'enable_dmfs=1'
```

➢ ***To disable VXLAN hardware offload:***

```
esxcli system module parameters set -m nmlx4_core -p 'enable_vxlan_offloads=0'
```

Except for the module parameters set above, the rest of VXLAN configuration is done in the ESX environment via VMware NSX manager. For additional NSX information, please refer to VMware documentation, see:
http://pubs.vmware.com/NSX-62/index.jsp?topic=%2Fcom.vmware.nsx.install.doc%2FGUID-D18A11DF-3D85-4B80-8713-D611648D43F4.html.

Additional information can be found at: http://dailyhypervisor.com/vmware-nsx-for-vsphere-6-1-step-by-step-installation/

4 Troubleshooting

You may be able to easily resolve the issues described in this section. If a problem persists and you are unable to resolve it yourself please contact your Mellanox representative or Mellanox Support at support@mellanox.com.

4.1 General Related Issues

Table 4 - General Related Issues

Issue	Cause	Solution
The system panics when it is booted with a failed adapter installed.	Malfunction hardware component	1. Remove the failed adapter. 2. Reboot the system.
Mellanox adapter is not identified as a PCI device.	PCI slot or adapter PCI connector dysfunctionality	1. Run `lspci`. 2. Reseat the adapter in its PCI slot or insert the adapter to a different PCI slot. If the PCI slot confirmed to be functional, the adapter should be replaced.
Mellanox adapters are not installed in the system.	Misidentification of the Mellanox adapter installed	Run the command below to identify the Mellanox adapter installed. `lspci | grep Mellanox'`

4.2 Ethernet Related Issues

Table 5 - Ethernet Related Issues

Issue	Cause	Solution
No link.	Mis-configuration of the switch port or using a cable not supporting link rate.	• Ensure the switch port is not down • Ensure the switch port rate is configured to the same rate as the adapter's port
No link with break-out cable.	Misuse of the break-out cable or misconfiguration of the switch's split ports	• Use supported ports on the switch with proper configuration. For further information, please refer to the MLNX_OS User Manual. • Make sure the QSFP break-out cable side is connected to the SwitchX.
Physical link fails to negotiate to maximum supported rate.	The adapter is running an outdated firmware.	Install the latest firmware on the adapter.

Table 5 - Ethernet Related Issues

Issue	Cause	Solution
Physical link fails to come up.	The cable is not connected to the port or the port on the other end of the cable is disabled.	Ensure that the cable is connected on both ends or use a known working cable

4.3 Installation Related Issues

Table 6 - Installation Related Issues

Issue	Cause	Solution
Driver installation fails.	The install script may fail for the following reasons: • Failed to uninstall the previous installation due to dependencies being used • The operating system is not supported	• Uninstall the previous driver before installing the new one • Use a supported operating system and kernel